Crucial stages of protein folding through a solvable model: predicting target sites for enzyme-inhibiting drugs.

Cristian Micheletti, Fabio Cecconi, Alessandro Flammini and Amos Maritan
International School for Advanced Studies (SISSA/ISAS), Trieste, ITALY
(Dated: February 1, 2008)

An exactly solvable model based on the topology of a protein native state is applied to identify bottlenecks and key-sites for the folding of HIV-1 Protease. The predicted sites are found to correlate well with clinical data on resistance to FDA-approved drugs. It has been observed that the effects of drug therapy are to induce multiple mutations on the protease. The sites where such mutations occur correlate well with those involved in folding bottlenecks identified through the deterministic procedure proposed in this study. The high statistical significance of the observed correlations suggests that the approach may be promisingly used in conjunction with traditional techniques to identify candidate locations for drug attacks.

INTRODUCTION

One of the open fundamental questions in molecular biology is how to predict the folded state of a protein from the knowledge of its sequence. Despite a large increase in available computing power in the past years, it has been impossible to answer this question by means of computer simulations of various degrees of complexity and detail. However an increasing amount of experimental [1, 2, 3, 4, 5] and theoretical results [6, 7, 8, 9, 10] supports the view that the folding of natural proteins into their native state is largely influenced by the native-state topology (for a brief review see [11]). Accordingly, the folding process is regarded as a well defined sequence of obligatory steps to be taken in order to reach the native state. Even if protein sequences have evolved to fold efficiently, the kinetics en-route to the native state might be hindered by the realization of particularly difficult (rate-limiting) steps, such as the formation of non-local amino acid interactions (contacts) that usually requires the overcoming of large entropy barriers. Some non-local native contacts are rather crucial for the folding process, because their formation helps establishing further native interactions and leads to a rapid progress along the folding pathway until another barrier is met. Their formation is associated to bottlenecks for the entire folding process. Strikingly, the amino acids involved in such crucial contacts are those for which the largest changes in the folding kinetics are observed in site-directed mutagenesis experiments [1], as first proven for CI2 and Barnase [6]. This suggests that protein sequences have been carefully optimised so to exploit the conformational entropy reduction accompanying the folding process [12] through the selection of the key amino acids. The number and importance of bottlenecks depends significantly on several factors. Among the most important are the contact order of the protein[8] and whether it folds in two or more stages [13].

In previous studies [14, 15], we have shown how the most delicate folding stages can be identified within a molecular dynamics approach, by monitoring the formation probability of native and non-native contacts from the unfolded to the native state. This can either be done as a function of time at a fixed temperature around the folding temperature or working at thermal equilibrium for a succession of decreasing temperatures (annealing). In principle, the two approaches need not to be equivalent but, for the quantities we have investigated, they give consistent results. Then, concerning the indentification of crucial contacts, one can safely concentrate on studying thermodynamic equilibrium at various temperatures. The main limitation of Molecular Dynamics (MD) and MonteCarlo (MC) simulations, especially for long protein chains, is that they are extremely time-demanding and plagued with statistical errors which can affect the predictions based on the study of the relative sensitivity of contact formation. Therefore it would be highly desirable to develop a suitable theoretical model, amenable to a deterministic (and computationally fast) treatment, thus resulting in a deeper understanding of the problem. Ideally, such a model should encompass all the "necessary ingredients" that usually are included in computer simulations: peptide chain constraints, effective interactions between residues, favourable monomeric positions, etc. In the following we describe a recently developed theoretical scheme [16], that, while being very simplified and approximate compared to other schemes based on MD or MC simulations, can be treated analytically, leading to expressions that can be evaluated exactly. The calculated quantities rival those obtained through more sophisticated but computationally demanding MC and MD techniques. The purpose of the present paper is to show how the model can be employed to yield helpful observables to identify the folding bottlenecks. In particular we apply the method to the HIV-1 Protease (HIV-1 PR), an enzyme which is crucially involved in the HIV infection [17]. In general, the accurate knowledge of bottlenecks has important pharmaceutical ramifications because their knowledge may be exploited in a rational drug design. Due to the large

amount of available clinical data, HIV-1 PR is a natural choice for a stringent test of our automated predictive scheme.

THEORY

The model we adopt builds on the importance of the native state topology in steering the folding process, that is in bringing into contact pairs of amino acids that are found in interaction in the native state. A primary quantity of interest that we shall calculate is the probability that a given native contact is established at a definite stage of the folding process. Probably, the oldest attempt to calculate such quantity dates back to Flory who tried to estimate the probability p_{ij} that two sites i and j in a long harmonic chain (the peptide) are in contact [18]. The approximation introduced by Flory was to neglect correlations between residues, which amounts to considering the chain embedded in a highly-dimensional space. As a result, the p_{ij}'s are a decreasing function of the sequence separation $|i-j|$. Clearly, this approximation is not apt to pinpoint the key folding sites, since it exploits the native topology at the simplest level; in fact it takes into account only the contact order of native interactions. The Flory approach, however, can be refined by incorporating correlations between the formation of pairs, triplets etc. of contacts [19, 20, 21]. Here we use a recently introduced energy function that allows to calculate the p_{ij}'s within a self-consistent analytic scheme. The strategy is similar in spirit to that of Go and Scheraga [22] where only the formation of native interactions is energetically rewarded and is common to all recent approaches which exploits the native state topology [6, 7, 8, 9, 10].

We describe the proteins by the coordinates $\mathbf{r}_i$ of the C_α atom of the i-th amino acids. The simplified energy functional for the chain of N residues is

$$H = \frac{KT}{2}\sum_{i=1}^{N-1}(\mathbf{r}_{i,i+1}-\mathbf{r}^0_{i,i+1})^2 + \frac{1}{2}\sum_{i\neq j}\Delta_{ij}[(\mathbf{r}_{ij}-\mathbf{r}^0_{ij})^2 - R^2]\theta_{ij} \tag{1}$$

where K is the strenght of the peptide bonds, assumed to be harmonic, and T is the absolute temperature in units of the Boltzmann constant.

The relative position between amino acid centroids is denoted by $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and the corresponding native positions are indicated with the superscript 0. Δ is the contact matrix, whose element Δ_{ij} is 1 if residues i and j are in contact in the native state (i.e. their C_α separation is below the cutoff $c = 6.5$ Å) and 0 otherwise. The matrix Δ_{ij} along with the set $\mathbf{r}^0_{ij}$ encodes the topology of the protein. The factor θ_{ij} has the form

$$\theta_{ij} = \Theta(R^2 - (\mathbf{r}_{ij}-\mathbf{r}^0_{ij})^2) \tag{2}$$

where $\Theta(x)$ is the unitary step function and R is a distance cutoff defining the range of the interaction between non consecutive amino acids. In standard off-lattice approaches, the interaction $V(d)$ between non-bonded amino acids at a distance d, is taken to be a square well potential, or some type of Lennard-Jones interaction. Our choice in Eq. (1) is a sort of "harmonic well" which, while being physically sound and viable, is suitable for a self-consistent treatment, as explained below. The location of the outer rim of the well is controlled by R, which can be set to a few Angstroms ($R = 3$ Å in the present study) to penalise conformations where the the separation of two residues differs too much from the native one. In the native state each θ_{ij} is close to 1 while in the denaturated state case θ_{ij} are usually negligible.

While the present form of the model does not accurately describe the effects of self-avoidance this does not lead to a qualitatively wrong behaviour in the highly-denatured ensemble (large T). The treatment of steric effects becomes progressively more accurate as temperature is lowered. In fact, the model guarantees that the native state is the true ground state and therefore protein conformations found at low temperature inherit the native self-avoidance. The connectedness of the chain, as well as its entropy, are captured in a simple but non-trivial manner. The most significant advantage of the model is that it can be used to explore the equilibrium thermodynamics without being hampered by inaccurate or sluggish dynamics.

Two limit cases of model 1 are worthy of notice. In the absence of any bias towards the target structure (i.e. when both Δ_{ij} and the $\{r^0\}$'s are removed) the model reduces to the standard gaussian polymer model whose behaviour is exactly known [18, 23]. Furthermore, the limit when $T \to 0$ (when all native contacts are established and the bonded-energy term fluctuations are negligible) the model reduces to the gaussian network model that has been introduced and used to study the near-native vibrational properties of several proteins [24, 25, 26, 27].

The thermodynamics of the model is fully determined by the partition function

$$\mathcal{Z}(T) = \int \prod_{i=1}^{N} d^3 r_i \exp(-H/T). \tag{3}$$

In the integral of equation (3) and in the following, it is always meant that translational invariance is always explicitly broken by fixing, for example, the centre of mass of the system (see Appendix).

The integral (3) is still hard to treat analytically, due to the presence of non-quadratic interactions in the last term of Hamiltonian (1). We thus perform a further, but non-trivial, simplification by replacing H with the variational hamiltonian H_0

$$H_0 = \frac{KT}{2}\sum_{i=1}^{N-1}(\mathbf{r}_{i,i+1} - \mathbf{r}^0_{i,i+1})^2 + \frac{1}{2}\sum_{i\neq j}\Delta_{ij}[(\mathbf{r}_{ij} - \mathbf{r}^0_{ij})^2 - R^2]p_{ij} \tag{4}$$

where the factors θ_{ij} are now subtituted by parameters p_{ij} independent of the coordinates. Due to its quadratic form, the model described by Eq. (4) can be solved with the standard techniques for Gaussian integrals. Such parameters have to be optimally determined so to ensure self-consistency:

$$p_{ij} = \langle\Theta((\mathbf{r}_{ij} - \mathbf{r}^0_{ij})^2 - R^2)\rangle_0. \tag{5}$$

The symbol $\langle ...\rangle_0$ indicates that the thermal averages are performed through the Hamiltonian H_0. Now in such self consistent approach the problem is fully solved and we can compute the resulting partition function from which we extract all the thermal properties and averages. In particular the logarithm of the partition funzion, $\ln(\mathcal{Z})$, has the following explicit expression:

$$\ln(\mathcal{Z}) = \frac{3(N-1)}{2}\ln(2\pi) - \frac{3}{2}\ln N - \frac{3}{2}\ \ln(\det{}'M) + \frac{R^2}{2T}\sum_{lm}\Delta_{lm}\ p_{lm} \tag{6}$$

where the matrix M is defined as:

$$M_{i,j} = \begin{cases} K(2-\delta_{i,1}-\delta_{i,N}) + 2\sum_l \Delta_{i,l}\, p_{i,l}/T & \text{for } i = j \\ \\ -2p_{i,j}\Delta_{i,j}/T - K\,[\delta_{i,j+1} + \delta_{i,j-1}] & \text{for } i \neq j\ . \end{cases} \tag{7}$$

and the prime in (6) denotes that the zero eigenvalue of M has to be omitted (see Appendix).

The quantities p_{ij} in Eq. (5) represent precisely the occurrence probability of a contact between residues i and j and indicate the frequency with which that native contact is established. At thermal equilibrium their dependence on temperature reflect the status of compactness of the protein molecule. For instance, well below the folding temperature, T_F, each $p_{ij}(T)$ is expected to assume a value close to unity, as all native contacts are already formed. Instead, for temperatures much larger than T_F, all $p_{ij}(T)$ tend to be very small, reflecting the low propensity of the protein to establish contacts.

Thermodynamics quantities can be easily derived from the p_{ij}'s. Another quantity necessary to characterize the folding transition is the specific heat, which exhibits one or more peaks in correspondence of significant structural rearrangements of the protein conformation. Since every energy change is mainly associated to the formation of native interactions, we address the question of which native contacts contribute mainly to the peak(s) of the specific heat. A clear answer to this question is readily found in the temperature behaviour of frequencies p_{ij}. Indeed, each $p_{ij}(T)$ exhibits a sigmoidal shape, and the modulus of its derivative develops a sharp maximum in correspondence of the point of inflection (crossover temperature). The importance of every native contact $i-j$ turns out to be characterized by the crossover temperature and the maximum slope of its p_{ij}, which can be regarded as an indicator of its degree of cooperativity. In fact, the most important contacts are those with high crossover temperature and associated high cooperativity.

This fact allows a complete identification and classification of the bottlenecks, because we are now able to indentify those contacts that are termodinamically relevant to peaks and shoulders of the specific heat.

APPLICATION TO HIV-1 PROTEASE

The human immunodeficiency virus (HIV) encodes a protease, HIV-1 PR, whose inhibition is crucial to prevent the maturation of infectious HIV particles [17]. The role of the Protease in the infection spreading is to act as "molecular scissor" cleaving inactive viral polyproteins into smaller, functional proteins. In the presence of protease inhibitors,

viral particles are unable to mature and are rapidly cleared. Extensive clinical trials have lead to the development of five HIV-1 PR inhibitors approved by the Food and Drug Administration (FDA): Saquinavir mesylate (SAQ), Ritonavir (RIT), Indinavir sulfate (IND), Nelfinavir mesylate (NLF) and Amprenavir (APR) [28]. Such drugs are particularly effective in short-term treatments, while their long-term efficacy is limited by resistance. Indeed mutants resistant to protease-inhibitors can emerge in vivo already after less than one year [17]. Table I summarises the list of HIV-1 PR known mutating sites causing drug resistance.

In an earlier work, the study of the near-native harmonic vibrations of the HIV-1 PR has shown that a number of sites that are paramount to the stability of the native enzyme are close to some of the residue of Table 1 [25]. The self-consistent scheme of eqn. 4 allows to extend this result by modelling the partially-folded ensemble at finite temperature.

In particular, we will be concerned in the characterization of such ensemble near the folding transition temperature. The motivation to do so stems from a recent study [14] where we have shown that such mutating amino acids correspond, with high statistical significance, to sites involved in the folding kinetic bottlenecks. The rationale for this finding is that the most effective drugs can be eluded only by mutations occurring in correspondence of the key sites. Due to the sensitivity of the folded native conformation to these sites, only fine-tuned mutations are allowed in correspondence of these sites. Such mutation have to result in a native-like enzymatic activity and in the avoidance of the drug action. These constraint act as a severe selective pressure on the mutated proteases that the HIV virus is able to express. As a result, the mutations that will ultimately cause drug-resistence are expected to occur in correspondence of the crucial sites. These residues are heavily influenced by the native topology, and hence should display little dependence on the particular (effective) drug to be eluded.

It is therefore our purpose to apply the scheme introduced in the previous section and identify the key residues within our topology-based scheme. The method, being completely analytic, is free from statistical uncertainty, common to all MC and MD simulation methods, or from difficulty (due to spatial restraints) to reach the target native state below the folding temperature.

RESULTS AND DISCUSSION

The structural model at the basis of our analysis is the free enzyme [17]. It is a homodimer with C2 symmetry, each subunit being composed by 99 residues (Fig. 1). Previous studies [14] have shown that geometrically important residue positions can be obtained considering a single monomer. Indeed the specific heat of the whole homodimer on decreasing the temperature shows a peak in correspondence of the folding of each sub-unit and then at lower temperature another peak signals the aggregation of the two sub-units. Thus, in the following, we will be concerned only with a single monomer. The specific heat is obtained through numeric differentiation of the average internal energy, which has the following explicit analytic expression in terms of the $p_{ij}(T)$'s and the quantities introduced before:

$$\langle E \rangle = \frac{3\,(N-1)\,T}{2} - \frac{R^2}{2} \sum_{ij} \Delta_{ij}\, p_{ij}(T)\ . \tag{8}$$

The study of Go and Scheraga [22] showed that systems described by energy-scoring-functions that reward the formation of native contacts display cooperative (all-or-none) folding transitions with an associated peak(s) in the specific heat. Consistently with these expectations, the specific heat calculated by differentiating Eq. (8) with respect to T shows a single peak, see Fig. 2, thus providing an unambiguous criterion for identifying the folding transition temperature, T_F. The width of the specific heat peak at the folding transition in Figure 2 is larger than the typical one found in experimental [13] and theoretical studies [29, 30]. It is possible to enhance the cooperativity of the transition by intervening on the actual value of K in Eq. (1); in fact, a decrease of K leads to sharper transitions. An alternative criterion for fixing the value of K is provided by its influence on the average amount of native structure that is formed at the native state. Since we are particularly interested in monitoring the progressive establishment of native contacts, we adopted this second possibility to set the value of K. In fact, by choosing $K = 1/15$ in (1), we ensure that, at T_F, the average fractional occupation of native contacts, q:

$$q = \frac{\sum'_{i,j} \Delta_{i,j}\ p_{i,j}}{\sum'_{i,j} \Delta_{i,j}} \tag{9}$$

is about 50 % (see Fig. 2), as established in several experiments and numerical studies. The primed summation symbol indicates that the sum is not carried out over consecutive pairs. The degree of native similarity, q is a useful overall indicator to monitor the progress towards the native state in a folding process [31, 32]. While the ultimate quantities of interest are the $p_{i,j}$'s, it is useful to consider an intermediate level of description and focus on the whole network of contacts that a given site takes part to. A natural order parameter is provided by the "average environment formation" [33, 34] which, for a generic site i is defined as:

$$P_i = \frac{\sum_j' \Delta_{i,j}\ p_{i,j}}{\sum_j' \Delta_{i,j}}\ . \tag{10}$$

P_i is a measure of the fraction of established native contacts the i-th residue partecipate to (clearly, P_i is defined only when the denominator of eqn. 10 is non-zero). The environment profiles for three different temperatures are shown in Fig. 3. The irregular behaviour of the profiles results from a complex interplay of the burial of the sites and the locality of their contacts. The hierarchical formation of secondary structures at high temperature is clearly visible. It is instructive to correlate the location of the sites known to cause resistance to drug treatments (see Table I) with the features of the profiles. In particular, several mutating sites responsible for drug resistance (see Table I) can be found in correspondence of the peaks of the environments (see in particular sites 20,63,71,77,84). The most precise way to identify the key residues is, however, through the analysis of the fractional occupation of native contacts and not through the environments, since they only carry averaged information. Typical $p_{i,j}$ curves as a function of temperatures are shown in Fig. 4.

As anticipated in section Theory, all $p_{i,j}$'s have monotonic sigmoidal shapes which mainly reflect the sequence separation, $|i-j|$ and the native burial of each of the residues. In general, each contact is established at a different crossover temperature and with different intensity [14]. The data relative to the frequencies of native-contact formation is conveniently summarised in the color-coded contact maps of Fig. 5. A bright red color is used to highlight those contacts with the largest crossover temperatures above T_F, see Fig. 5a, or highest intensity in Fig. 5b. Both these intuitive notions can be used to identify the key folding contacts. The inspection of Fig. 5 reveals that several kinetic bottlenecks (red regions) are located three-four contacts downstream the three β-turns in HIV-1 PR. In addition, the formation of contacts around residues 84 and 30, despite being so far away along the sequence, appears to be a crucial folding stage since it allows the collapse of the individual secondary structure motifs. It is striking that these results make an excellent parallel with those of Ref. [14], where long and delicate MD simulations of the unfolding/refolding of HIV-1 PR were carried out using a much more sophisticated energy-scoring function. This provides a cross validation for the robustness of the results obtained both in the stochastic and the present, analytic, scheme. The emphasis is on the exactness of the present approach that allows to determine easily the $p_{i,j}$'s with an arbitrary accuracy. The absence of stochastic noise allows to compile Table II which shows the top contacts ranked according to crossover temperature and intensity. Sites that are known to cause drug resistance through mutations are highlighted in boldface. It is apparent that a high fraction of the top key folding contacts do, indeed, contain key mutating sites. To test the significance of such matches we compare the number of marked mutating sites contained in each column of Table II with the number of those contained in a randomly compiled table. We expect a random list of t elements extracted among N, m of which are marked, to contain an average of tm/N marked elements with a square deviation of $tm(N-m)(N-t)/(N^2(N-1))$. For the case of HIV-1 PR the total number of contacts (excluding consecutive residues) within a cutoff radius of 6.5 Å is $N = 180$ and the number of those which include at least one known mutating site is $m = 60$. An analysis of the contacts of Table II (selected according to crossover temperature or cooperativity of formation) shows that the number of matches observed among the top sites tipically exceeds that expected from a random choice by a standard deviation (the precise amount depend of how many top ranking contacts are considered. An alternative and apparently more stringent approach is to identify independent groups of highly correlated contacts, and then search for the key residues in each group. To a first approximation, the correlated sets of interacting pairs may be identified with the clusters in the contact map. This leads to define six main groups, the three β-sheets, the helix and the two sets of long-range contacts, around contacts 14-60 and 23-84, respectively (see Fig. 5). The four contacts in each group with the highest intensity of formation above T_F are summarised in Table III. Out of the 24 contacts, 12 of them involve a key site, which is two standard deviations away from the number of matches expected on a random basis (7.9 ± 2.1). Again, this testifies both the reliability of the general scheme followed here and also its robustness in the different possible implementations.

Interestingly, the results of Table III account better than those of Table II for the heterogeneous location of the key folding sites. The emerging conclusion is that a complete description of the crucial contacts can be obtained only by monitoring all the key stages of the folding process. In standard MC and MD simulations of protein unfolding/refolding, it is the simulated dynamics that reveals which, and how many, delicate stages exists. In the present

approach, the folding process is characterised analytically, thus the complete set of folding bottlenecks follows from the study of distinct groups of interrelated contacts.

Finally, we remark that the determination of the key contacts does not uniquely provide the key folding sites, since two sites are involved in each pairwise contact. This ambiguity can, in several cases, be resolved either by selecting those sites that take part in several crucial contacts, or by examining their distribution on the three-dimensional native structure for clues that may help breaking the ambiguity.

CONCLUSIONS

We have used an analytical technique to study and characterize the folding process of globular proteins. This deterministic method allows the automated identification of contacts involved in folding rate-limiting steps. As a result, the whole folding process is particularly sensitive to mutations occuring at sites involved in such crucial contacts. We test our scheme and its usefulness in pinpointing the crucial sites by applying it to HIV-1 protease. For this enzyme, extensive clinical trials have allowed the identification of several sites involved in drug-resistance mutations. Such sites have a meaningful overlap with the key folding sites predicted by our scheme with a modest computational effort compared to more sophisticated stochastic simulations techniques. This indicates that the available inhibiting drugs are quite effective since they can be eluded only by mutations of the (sensitive) key sites of the protease.

The proposed approach to identify the crucial residues is quite general and ought to be useful to identify the kinetic bottlenecks of other viral enzymes of pharmaceutical interest, thus aiding the development of novel effective inhibitors.

We expect to focus our future efforts on improving the present approach by taking into account the propensities of different amino acids to form contacting pairs. This limitation can be overcome by introducing physically viable (attractive) pairwise interactions [35, 36, 37, 38, 39]. In the present approach this possibility was deliberately avoided to highlight the influence of the native state topology alone on the kinetic bottlenecks, irrespective of the different chemical nature and strength of the effective amino acid interactions. We expect that the inclusion of such effects, while not distorting the overall picture presented here, may change the relative strength of spatially-close contacts. This may improve the agreement between Table I and tables II-III by resolving those cases were a site adjacent to a mutating one is selected.

We are indebted to Paolo Carloni for several illuminating discussions and for having stimulated the present work. This work was supported by INFM, Murst Cofin2001.

APPENDIX

In this appendix we discuss how the translational invariance of a quadratic energy scoring function can be explicitly broken by fixing the center of mass of the system in the origin. The constrained partition function is written as:

$$\mathcal{Z} = \int \prod_{i=1}^{N} d^3x_i \; e^{-1/2 \sum_{i,j} \mathbf{x}_i \, A_{ij} \, \mathbf{x}_j} \; \delta^3(\sum_i \mathbf{x}_i) \tag{11}$$

where the matrix A incorporates the quadratic dependence of H_0 in eqn (4) from the space co-ordinates (and also includes the $1/T$ factor to yield the usual Boltzmann weight). The translational invariance of H_0 implies that A satisfies the property: $\sum_j A_{ij} = 0$, which amounts to say that the uniform vector, $\mathbf{v}_1 \equiv N^{-\frac{1}{2}} (1, 1, 1, 1..., 1)$ is an eigenvector of A with eigenvalue $\lambda_1 = 0$. We assume that H_0 is invariant only for the simultaneous translation of all the coordinates, $\{\mathbf{x}_i\}$. In this case all other eigenvalues, $\{\lambda_{i>1}\}$ are strictly positive and the corresponding eigenvectors $\mathbf{v}_{i>1}$ are all orthogonal to the zero mode $\mathbf{v}_1$.

By rewriting the Dirac-delta constraint as

$$\delta^3(\mathbf{z}) = \lim_{c \to \infty} \left(\frac{c}{2\pi} \right)^{\frac{3}{2}} e^{-c \; \mathbf{z} \cdot \mathbf{z}/2} \tag{12}$$

the partition function takes on the form $\mathcal{Z} = \lim_{c \to \infty} \mathcal{Z}_c$, where

$$\mathcal{Z}_c = \left(\frac{c}{2\pi} \right)^{\frac{3}{2}} \int \prod_{i=1}^{N} d^3x_i \; e^{-1/2 \sum_{i,j} \mathbf{x}_i \, A'_{ij} \, \mathbf{x}_j} \; , \tag{13}$$

where $A'_{ij} = A_{ij} + c$. It is straightforward to see that A' admits the same eigenvectors of A. Only the zero mode eigenvalue will change from zero to $c\,N$, while all others will be unmodified. Upon performing the Gaussian integrations in $\mathcal{Z}_c$ we obtain:

$$\begin{aligned} \mathcal{Z}_c &= \left(\frac{c}{2\pi}\right)^{\frac{3}{2}} \left(\frac{2\pi}{cN}\right)^{\frac{3}{2}} \prod_{i=2}^{N} \left(\frac{2\pi}{\lambda_i}\right)^{\frac{3}{2}} \\ &= \frac{1}{N^{3/2}}\,(2\pi)^{\frac{3(N-1)}{2}} \prod_{i=2}^{N} {\lambda_i}^{-\frac{3}{2}}\ . \end{aligned} \tag{14}$$

This shows that $\mathcal{Z}_c$ is effectively independent of c and, therefore, the partition function $\mathcal{Z}$ simplifies to

$$\mathcal{Z} = N^{-\frac{3}{2}}\ (2\pi)^{\frac{3(N-1)}{2}}\ (\det{}'A)^{-\frac{3}{2}}\ , \tag{15}$$

where the prime denotes that the determinant is calculated omitting the zero mode eigenvalue.

[1] A. R. Fersht, Proc. Natl. Acad. Sci. USA **92**, 10869 (1995).
[2] J. C. Martinez and L. Serrano, Nature Struct. Biol. **6**, 1010 (1999).
[3] D. S. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker, Nature Struct. Biol. **6**, 1016 (1998).
[4] F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson, Nature Struct. Biol. **6**, 1005 (1999).
[5] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[6] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, Phys. Rev. Lett. **82**, 3372 (1999).
[7] A. Maritan, C. Micheletti, and J. R. Banavar, Phys. Rev. Lett. **84**, 3009 (2000).
[8] E. Alm and D. Baker, Proc. Natl. Acad. Sci. USA **96**, 11305 (1999).
[9] C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[10] T. X. Hoang and M. Cieplak, J. Chem. Phys. **113**, 8319 (2000).
[11] D. A. Baker, Nature **405**, 39 (2000).
[12] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[13] S. E. Jackson, Folding and Design **3**, R81 (1998).
[14] F. Cecconi, C. Micheletti, P. Carloni, and A. Maritan, Proteins: Structure Function and Genetics **43**, 365 (2001).
[15] G. Settanni, C. Cattaneo, and A. Maritan, Biophys. J. **80**, 2935 (2001).
[16] C. Micheletti, J. Banavar, and A. Maritan, Phys. Rev. Lett. **87**, DOI:088102 (2001).
[17] J. H. Condra et al., Nature **374**, 569 (1995).
[18] P. J. Flory, J. Am. Chem. Soc **78**, 5222 (1956).
[19] H. S. Chan and K. A. Dill, J. Chem. Phys. **92**, 3118 (1990).
[20] D. A. Debe and W. A. Goddard III, J. Mol. Biol. **294**, 619 (1999).
[21] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA **92**, 1277 (1995).
[22] N. Go and H. A. Scheraga, Macromolecules **9**, 535 (1976).
[23] A. Kloczkowski and R. L. Jernigan, Comp. Theor. Pol. Sci. **9**, 285 (1999).
[24] I. Bahar, A. R. Atilgan, and B. Erman, Folding and Design **2**, 173 (1997).
[25] I. Bahar, B. Erman, R. L. Jernigan, A. R. Atilgan, and D. G. Covell, Journal of Molecular Biology **285**, 1023 (1999).
[26] O. Keskin, I. Bahar, and R. L. Jernigan, Biophysical Journal **78**, 2093 (2000).
[27] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Biophysical Journal **80**, 505 (2001).
[28] P. J. Ala, E. E. Huston, R. M. Klabe, P. K. Jadhav, P. Y. S. Lam, and C. H. Chang, Biochemistry **37**, 15042 (1998).
[29] H. Kaya and H. S. Chan, Phys. Rev. Lett. **85**, 4823 (2000).
[30] H. Kaya and H. S. Chan, Proteins: Structure Function and Genetics **43**, 523 (2001).
[31] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA **90**, 6369 (1993).
[32] A. Sali, E. Shakhnovich, and M. Karplus, Nature **369**, 248 (1994).
[33] O. V. Galzitskaya and A. V. Finkelstein, Proc. Natl. Acad. Sci. USA **96**, 11299 (1999).
[34] T. Lazaridis and M. Karplus, Science **278**, 1928 (1997).
[35] F. Seno, C. Micheletti, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **81**, 2172 (1998).
[36] M. J. Sippl, Curr. Opin. Struct. Biol. **5**, 229 (1995).
[37] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1999).
[38] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).
[39] C. Micheletti, F. Seno, J. R. Banavar, and A. Maritan, Proteins: Structure Function and Genetics **42**, 422 (2001).
[40] A. Molla et al., Nat. Med. **2**, 760 (1996).

[41] M. Markowitz et al., J. Virol. **69**, 701 (1995).
[42] A. K. Patick et al., Antimicrob. Agents Chemother. **40**, 292 (1996).
[43] M. Tisdale et al., Antimicrob. Agents Chemother. **39**, 1704 (1995).
[44] H. Jacobsen et al., J. Infect. Dis. **173**, 1379 (1996).
[45] P. Reddy and J. Ross, Formulary **34**, 567 (1999).

Drug	Point Mutations
RTN [40, 41]	20,33,35,36,46,54,63,71,82,84,90
NLF [42]	30,46,63,71,77,84,
IND [17, 43]	10,32,46,63,71,82,84
SQV [17, 43, 44]	10,46,48,63,71,82,84,90
APR [45]	46,63,82,84

TABLE I: Mutations in the protease associated with FDA-approved drug resistance [28].

Crossover Temperature	Cooperativity
25 - 86	14 - 66
28 - 86	14 - 64
58 - 76	**10** - 23
58 - **77**	14 - 65
57 - **77**	13 - 66
13 - 66	12 - 66
30 - 86	87 - 91
32 - **84**	13 - 65
32 - 76	23 - **84**
29 - 86	**10** - 22
31 - **84**	56 - **77**
23 - **84**	57 - **77**
14 - 66	23 - 83
25 - 85	22 - **84**
14 - 65	57 - 78
45 - 56	86 - 89
89 - 91	34 - 78
13 - 65	58 - **77**
87 - 89	**30** - 88
84 - 86	**32** - 75
56 - 58	**32** - 76
25 - **84**	31 - 76
86 - 88	42 - 58
64 - **71**	**90** - 94
57 - 76	87 - **90**

TABLE II: The top contacts ranked according to the crossover temperature (first column) and cooperativity of formation above T_F (second column)

Bottlenekcs	Key Contacts
β_1	**10** - 23
β_1	**10** - 22
β_1	14 - **20**
β_1	12 - **20**
β_2	42 - 58
β_2	45 - 58
β_2	43 - 58
β_2	43 - 57
β_3	56 - **77**
β_3	57 - **77**
β_3	58 - **77**
β_3	57 - 76
Other1	14 - 66
Other1	14 - 64
Other1	14 - 65
Other1	13 - 66
Other2	23 - **84**
Other2	23 - 83
Other2	22 - **84**
Other2	**30** - 88
Helix	87 - 91
Helix	86 - 89
Helix	**90** - 94
Helix	87 - **90**

TABLE III: The four contacts with the highest cooperativity of formation above T_F for each of the six clusters of the contact map.

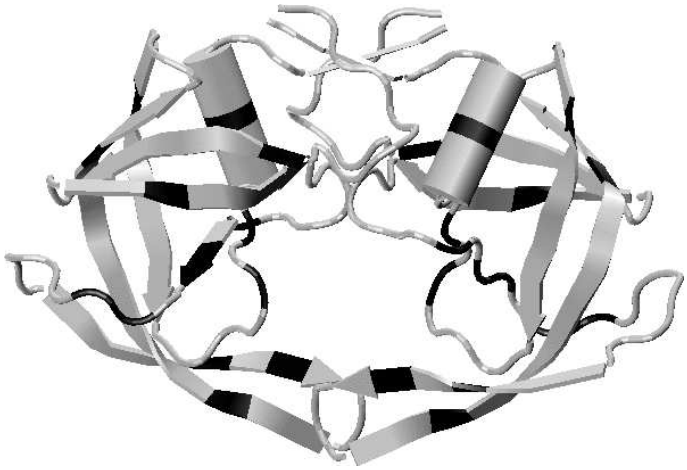

FIG. 1: Structure of HIV-1 PR dimer [17]. The highlighted locations indicate residues where mutations causing drug-resistance are observed.

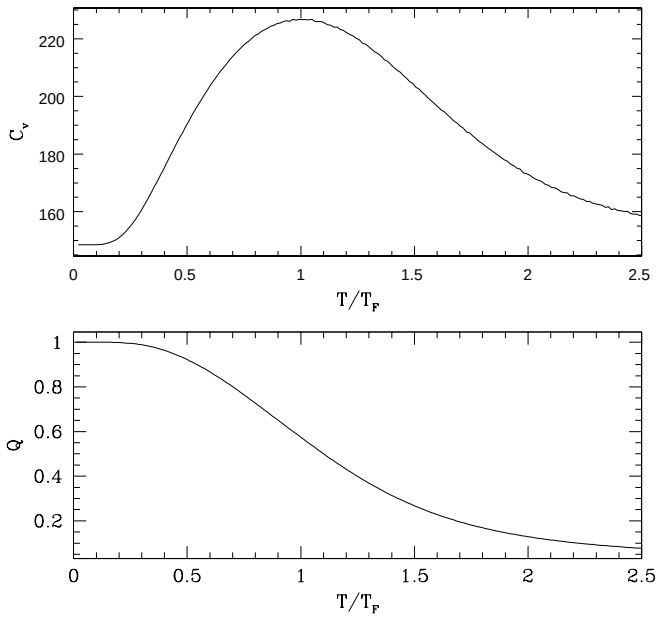

FIG. 2: Specific heat and overlap of a monomer of the HIV-1 PR. The temperature is scaled with the temperature T_F where the specific heat peak occurs

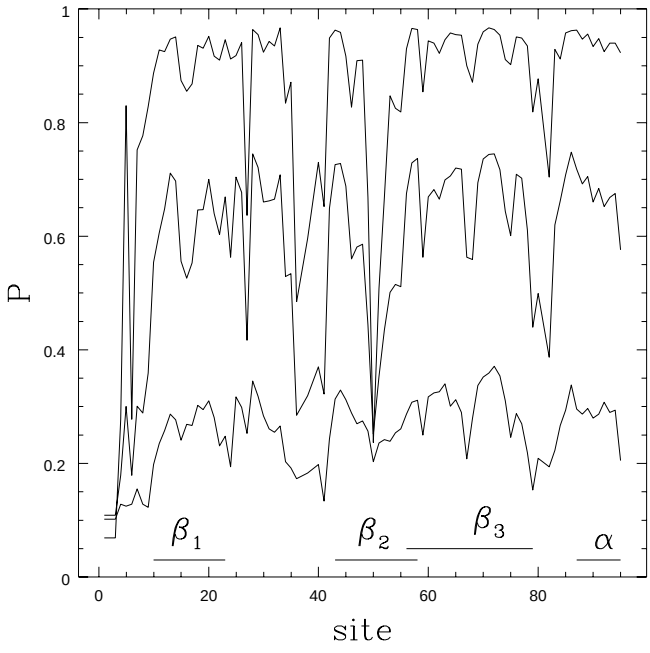

FIG. 3: Plot of P_i, the degree to which amino acid i is in a native-like conformation, versus i. In ascending order the curves are calculated at T/TF= 1.5, 1.0 and 0.5. The bar at the bottom shows the secondary structure associated with amino acid i.

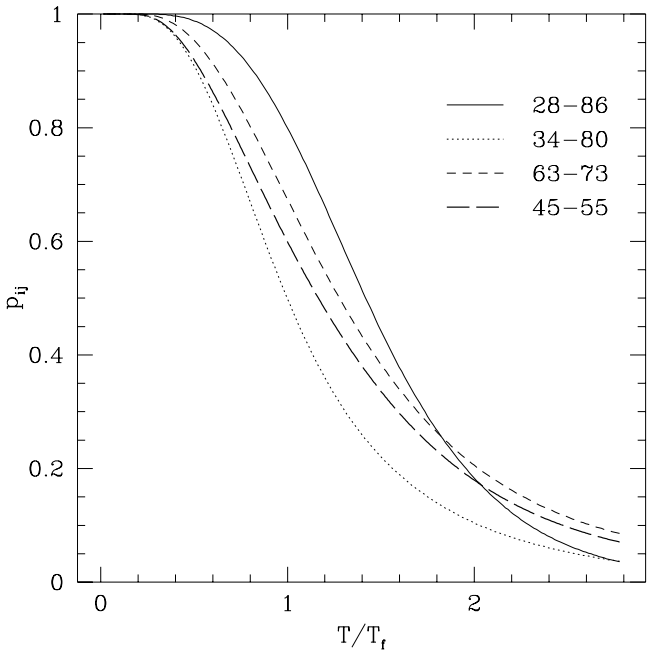

FIG. 4: Typical behaviour of contact probabilities, $p_{i,j}$ versus T/T_F for four native contacts involving pair of sites with different sequence separation and degree of native burial.

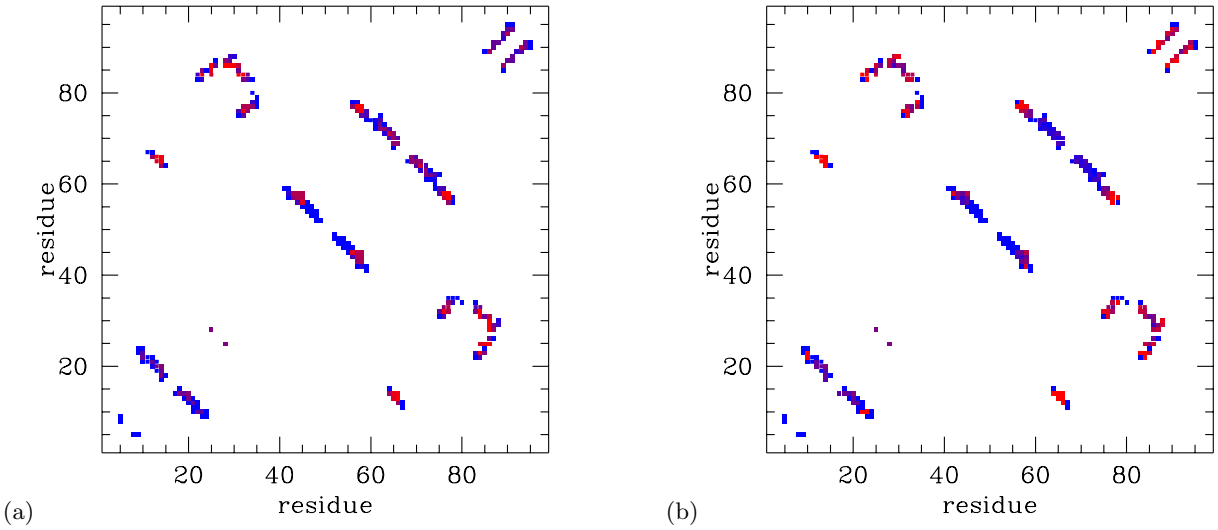

FIG. 5: Color-coded contact map of HIV-1 PR monomer. (a) Contacts with a large [small] crossover temperature are shown in red [blue].(b) Contacts with a large [small] cooperativity of formation above T_F are shown in red [blue].